DOCUMENT RESUME

ED 254 549 TM 850 158

AUTHOR Micceri, Theodore
TITLE Establishing the Reliability of the Florida Performance Measurement System's Research Based Observation Instrument.
PUB DATE Apr 84
NOTE 46p.; Paper presented at the Annual Meeting of the American Educational Research Association (68th, New Orleans, LA, April 23-27, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Analysis of Variance; *Classroom Observation Techniques; Elementary Secondary Education; *Evaluation Methods; Factor Structure; Generalizability Theory; Interrater Reliability; Teacher Behavior; *Teacher Evaluation; *Test Construction; *Test Reliability; Test Validity; Videotape Recordings
IDENTIFIERS *Florida Performance Measurement System

ABSTRACT

This paper investigates the reliability of the Florida Performance Measurement Systems' Summative Observation instrument. Developed for the Florida Beginning Teacher Evaluation Program, it provides behavioral ratings for teachers in a classroom setting. Data came from ratings of videotapes of nine teachers conducting actual lessons by nine teams of trained observers. Analysis of variance produced three estimates of reliability for each scale and subscale: (1) discriminant (across teachers); (2) stability (over time); and (3) interrater (among raters). Results indicate that the instrument appears sufficiently reliable to conduct classroom observations if ratings by at least two different observers are averaged to produce scores. Effective (positive) indicators of teacher behavior appear to be more reliably observed than ineffective (negative) indicators. Two domains--Management of Student Conduct, and Communication: Verbal and Nonverbal--appear too intercorrelated with the other domains for discrete reliable estimation of specific behaviors. Future research on this instrument should include validation, rater certification, norming and frame factors. Appendices contain: (1) background information on the knowledge base and the Florida teacher competencies; (2) indicators of the summative instrument; and (3) computation of reliability estimates. (BS)

ESTABLISHING THE RELIABILITY OF THE FLORIDA PERFORMANCE MEASUREMENT SYSTEM'S RESEARCH BASED OBSERVATION INSTRUMENT

by

Theodore Micceri

Paper presented at the AERA Annual Meeting
New Orleans, Louisiana -- April/1984

Summary

The study conducted on the reliability of the Florida Performance Measurement System's (FPMS) Summative Observation instrument supports the following conclusions:

* The instrument appears sufficiently reliable to conduct classroom observations if ratings by at least two different observers are averaged to produce scores.

* Effective (positive) indicators of teacher behavior appear to be more amenable to reliable observation than Ineffective (negative) indicators.

* Two domains, 2 - Management of Student Conduct, and 5 - Communication: Verbal and Nonverbal, appear too intercorrelated with the other domains for discrete, reliable estimation of specific behaviors. These domains should probably be used more as indicators than as judgement tools.

Acknowledgements

This paper is an attempt to report the work of numerous members of tne Florida Educational Community. The Florida Performance Measurement System has been in development for several years supported largely by volunteer efforts.

A core group consisting of B. Othanel Smith, Donovan Peterson, Jean Borg and Betty Fry have been the primary researchers and instrument developers throughout tne projects history. Robert Soar of the University of Florida, Donald Medley of the University of Virginia and Joseph Mazur of the University of South Florida were major contributors to the design of the reliability study conducted on the Summative Observation Instrument. In addition to these, Dave Berliner of the University of Arizona, and N.L. Gage of Stanford contributed to the content validation of the Domain Documents.

Donovan Peterson, B. Othanel Smith, Donald Medley, and Garfield Wilson provided contributions to the final form of this paper.

CONTENTS

LIST OF TABLES

Introduction

Purpose

The following study was conducted to support the reliability of an observation instrument developed to provide behavioral ratings for teachers in a classroom setting. Such ratings, if they prove feasible, may be used for problem identification, remediation and evaluation.

Classroom observations appear a necessity for evaluating teacher performance, since student outcome variables have not proven useful for this purpose. Observation techniques and instruments, however, are notoriously unreliable; suffering from such problems as rater inconsistency, lack of objectivity, unclear item definitions and changes over time in both observers and subjects. For these reasons, extensive and rigorous tests of reliability are recommended prior to the use of any observation instrument.

An observation instrument was developed for the Florida Beginning Teacher Evaluation Program, and tests of reliability were conducted to answer the question: "How justified are the researchers in generalizing the reliability estimates from this study to situations other than the one in which these estimates were obtained." The following three types of reliability were investigated to answer this question using a Three Way ANOVA model (Medley, 1982, Medley and Mitzel, 1963, Cronbach, 1972, Shrout, 1979):

1. DISCRIMINANT - consistency over subjects,

2. STABILITY - consistency over time, and

3. INTERRATER - consistency among raters.

Adequate outcomes within these three areas of reliability allow for generalization of the results to various teachers, observers, and teaching situations.

Instrument Development

The overall purpose of this research was to identify key elements that relate positively to student achievement. Basing their work on Florida legislation (section 231.29), a team of four education specialists conducted an extensive search of the process-product research literature and identified four observable domains of teacher behavior that consistently correlate with student achievement and also appear amenable to specification on an observation instrument.

DOMAIN 2	Management of Student Conduct
DOMAIN 3	Instructional Organization and Development
DOMAIN 4	Presentation of Subject Matter
DOMAIN 5	Communication: Verbal and Nonverbal

Their review indicated that not only were positive teacher behaviors associated with student achievement, but also, that specific and related negative behaviors appear to correlate negatively with achievement. Positive behaviors were termed Effective, and negative, Ineffective. Following content validation and pilot testing, the final version of a summative instrument contained twenty indicators of effective and twenty indicators of ineffective teacher behaviors. The indicators were couched in behavioral terminology as much as possible in order to reduce coding ambiguity. Each of the four domains identified above was represented in the instrument, however, not proportionally. The final form of this instrument contained:

DOMAIN 2	----	2 items
DOMAIN 3	----	11 items
DOMAIN 4	----	4 items
DOMAIN 5	----	3 items

An extensive study combining the expertise of university personnel, school district personnel and practicing teachers was conducted throughout the state of Florida during 1982-83 to clarify, content validate and test the reliability of this summative instrument.

Validation Procedures

Content Validity. The content validity of this instrument was supported by:

1. multiple independent sources for item development,

2. use of only research based indicators of teacher effectiveness for item development (indicators associated in the research with student achievement measures),

3. criticism and suggestions from knowledgeable persons external to the development, and

4. input from nationally known experts in the fields of teacher effectiveness, educational research and observation instruments.

Concurrent Validity. In an attempt to compare ratings obtained using the scaled scores from the summative instrument with ratings of knowledgeable persons regarding teacher behavior, two such people rated each of the teachers in the reliability study on a scale from 1 (inadequate) to 4 (excellent) for each of the domains, and for each of their lessons as a whole. Due to the low level of measurement embodied in this rating, a Spearman rank order correlation was run between these ratings and the standardized scores obtained during the reliablity study for each tape. There was a significant positive relationship between the ratings of these experts and the total scores for each tape on the summative instrument ($r = .55$). However, there was no significant relationship between any of the expert's domain subscale ratings and those of the instrument. This indicates either that the scale is more precise in evaluating specific domains, or that individual domains lack sufficient items to allow for specific evaluation using this instrument. Until this is resolved, only the total score should be used for decision-oriented evaluation. More study must be conducted before a reasonable understanding of the phenomenon may be proposed.

Method

Design

This study was designed to produce three intraclass correlation estimates of reliability derived from a three way Analysis of Variance (Medley 1982). The following main effect sources of variance are identified using ANOVA:

1. TEACHERS - variance among nine different teachers,

2. LESSONS - variance between two separate lessons taught by each of the nine teachers, and

3. RATERS - variance among nine teams of raters (observers).

These three main effects combined with various interaction effects are used to produce estimates of the following forms of reliability:

1. DISCRIMINANT - The consistency with which a test differentiates between different subjects (teachers) on a specific scale.

   If the instrument does not reliably discriminate among teachers having different behaviors, it cannot be used to evaluate levels of behavior.

2. INTERRATER - The consistency with which different raters score the same behavior exhibited by the same subject (teacher).

   If different raters do not produce consistent scores for the same teacher on the same lesson, one can assume either, that the items comprising the instrument are ambiguous, or that the raters are not adequately trained.

3. STABILITY - The consistency with which a specific subject (teacher) exhibits the same or similar behavior at two different times.

   Teachers were observed while teaching two lessons different in content, but similar in format. The teaching behaiors should not be altered substantially

by changing the content, as long as the lesson format remains the same. Variance accounted for by the interaction of raters and lessons, as well as that resulting from the teacher and lesson interaction provide an estimate of the stability of the teaching act over time.

Nine teachers were each observed teaching two lessons (a total of 18 lessons) by nine teams of raters. Eighteen video tapes of actual teacher's classroom behaviors were created, observed by the raters, and scaled into five separate scores: One total summative score and four subscales, one for each domain included in the summative instrument.

Scaling

In order to develop summative scores capable of rating teachers from HI to LOW on behaviors, a total scale based upon normalized scores for each individual item was created (see DATA TRANSFORMATION for transformation procedures applied). Each item on the instrument was standardized to a mean of 5, so that the mean total score was 100 (20 items times an average score of 5). One total scale for the 20 positive items and one for the 20 negative items was created. In addition, subscales were developed for each domain contained in the instrument. As a result of factor analysis conducted upon a preliminary version of this instrument, two items (#1 - Begins Instruction promptly, and #11 - Circulates and assists students) were transferred from the Domain 3 subscale to the Domain 2 subscale. Thus, the final version of Domain 2 contains four items instead of two, and Domain 3 includes nine items instead of eleven. The scores in this report consist of the summed normalized scores produced by teams of three observers for the following scales (see Appendix B for item content):

TABLE 1

Item Content of Scales

SCALE NAME	NUMBER OF ITEMS	ITEMS INCLUDED
TOTAL	20 items	1 thru 20
Domain 2	4 items	1,11,19,20
Domain 3	9 items	2 thru 10
Domain 4	4 items	12 thru 15
Domain 5	3 items	16 thru 18

Scaling Concerns. One major concern while developing scales for this instrument is the unusual nature of items 19 and 20 in Domain 2.

1. ITEM 19 -- stops misconduct

2. ITEM 20 -- maintains instructional momentum

Since these items deal specifically with control of behavior, their scale points have different meanings than other items. For the other 18 effective indicators, the more items marked, the higher a teacher's score. For items 19 and 20, however, the best possible score would be zero, as this would indicate a class under perfect control; requiring no teacher intervention to maintain momentum. The second best score would show only effective behaviors, and of course, the worst, only ineffective. This could cause some problems when interpreting a summative total score on the instrument. In practice, however, over several pilot tests, no teachers in the lower levels (below high school), exhibited zero behaviors for these items. For all practical purposes, the scales may therefore be considered uniform across all items, at least at the present time. The applicability of Domain 4 (Presentation of Subject Matter) to all levels of teaching poses another possible scaling problem. Some hypothesize that the use of this domain will vary substantially across grade levels, appearing only rarely in the

early grades, and becoming successively more common as grade level rises. It will be necessary to obtain substantially more data than is currently available to make decisive statements regarding this question.

Team Scores. Team scores were created by taking the mean transformed item scores for each team as the unit of analysis. These team item scores were then summed separately for the total instrument and each domain. Separate three way independent Analyses of Variance were conducted on each of the five resulting scales (total instrument, Domain 2, Domain 3, Domain 4, and Domain 5).

Reliability Estimates

Within reliability theory, the generally accepted definition of an obtained score is as follows:

$$O = T + E$$

where O = Obtained score
T = True score
E = Error score.

Reliability estimates are generally formulated in the following fashion using the elements of the preceding model:

$$r = 1 - (E/O)$$

This estimate separates the True score variance from the Error score variance present in the Obtained score (Nunnaly, 1978). The resulting reliability estimates are limited by the number of sources of variance identified by the model in use. From the sums of squares generated by the independent ANOVA, it is possible, using intraclass correlations, to create estimates of reliability that do not rest upon the rigorous set of assumptions underlying the F or T statistics. (Lindeman 1978, Medley and Mitzel 1963, Cronbach 1972, Guilford and Fructer, 1973).

This model uses the variance estimates produced by a three way ANOVA to identify souces of error across teachers, between lessons and among raters, as well as the interactions of these components. The variance estimates are then entered into an appropriate formula generated by classical reliability theory to obtain three separate and quite different estimates of reliability based on a single error term. Each of these estimates provides information regarding the generalizability of the results from this study.

TABLE 2

Independent ANOVA Sources for Sums of Squares

MAIN EFFECTS	NUMBER	df
(SST) Teachers	N	N-1
(SSR) Raters	K	K-1
(SSL) Lessons	D	D-1
INTERACTION EFFECTS		
(SSTXL) Teachers with Lessons		(K-1)(D-1)
(SSTXR) Teachers with Raters		(N-1)(K-1)
(SSRXL) Raters with Lessons		(N-1)(D-1)
RESIDUAL EFFECTS -- ERROR		
(SSRXLXT) Raters with Teachers with Lessons		(N-1)(K-1)(D-1)

Within the analysis conducted, error is considered that variance which is either unexplained or unaccountable.

TABLE 3

Resulting Reliability Estimates

TRUE SCORE = SST - SSTXL - SSTXR + SSRXTXL

ERROR SCORE = SSTXL + SSTXR + SSRXTXL

OBTAINED SCORE = TRUE SCORE + ERROR SCORE

ERROR for Stability = ERROR(ST)= SSTXL + SSRXTXL

ERROR for Rater Consistency = ERROR(R)= SSRXT + SSRXTXL

1. DISCRIMINANT (between teachers) = 1 - ERROR / OBTAINED
2. STABILITY (over time) = 1 - ERROR(ST) / OBTAINED
3. INTER-TEAM (among observers) = 1 - ERROR(R) / OBTAINED

Subjects

Raters. Forty two raters (observers) were trained in four Florida counties (Lee, Hillsboro, Pasco and Pinellas). The observers were volunteers, most of whom occupy administrative or supervisory positions in their county school districts. Each received five weeks of training (3-4 hours per week) prior to conducting observations for this study. Of the 42 total, 27 viewed at least 16 or more of the 18 video tapes. In an attempt to stabilize the observation scores, these 27 observers were randomly divided into teams of three observers each, and mean scores for teams were used as the unit of analysis.

Teachers. Nine teachers from Hillsboro and Orange counties were video taped while conducting actual lessons in a classroom setting. Teachers of various quality, styles, experience, grade levels and subject specialties were included in the study in an attempt to avoid possible bias in teaching style that might result from voluntary participation. Observers confirmed that a variety of teacher styles and quality were present in this study.

RESULTS

Responses

Seven hundred sixty eight observations were obtained from 42 observers of the 18 video tapes. Analysis was limited to 27 observers for whom at least 16 observations were available. This allowed for nine teams of three randomly assigned observers.

Scaling

Data Transformations. The item distributions in this study, as is typical for observation instruments, were characterized by extreme non-normality, with skews ranging as high as 5.2. For this reason, area transformations (Soar 1982) were conducted on the data to normalize the distributions. This was accomplished within the Statistical Analysis System (SAS) by first transforming each score to a percentile rank, to eliminate the extended nature of the tails; then standardizing the rank transformed data to further normalize the distributions. Both percentile and normal transformations were conducted using the SAS Rank subprogram. This resulted in individual item distributions more closely corresponding to the theoretical Gaussian (normal), at least with regard to skew, kurtosis, and the relationship between the standard deviation and the semi-interquartile range. Team scores were created from the transformed item scores. The three individual scores for each team were averaged to create a mean team score on each item. These item scores were then summed into scale scores for the total instrument and each domain separately. See Table 1 for specific scale composition.

Data Verification. Data obtained from this study were entered onto disk packs and accessed using the IBM 370 at the University of South Florida (Tampa, Fl.). All data were independently duplicated onto a second data set. The two data sets were compared; all discrepencies were referred back to the original observation instruments, and both data sets were corrected.

Analyses and Specific Results

Reliability of Effective Scales. Separate three way independent Analyses of Variance were conducted on each of the five resulting scales (total instrument, domain 2, domain 3, domain 4, and domain 5). This produced three estimates of reliability for each scale and subscale:

1. DISCRIMINANT - across teachers,
2. STABILITY - over time, and
3. INTERRATER - among raters

Table 4 gives the results obtained for the five scales identifying effective indicators of teacher behavior for the entire study (nine teams, nine teachers, two lessons). As one can see, these reliability estimates for the entire study are exceptionally high, with only domains 2 and 5 having estimates below .88, and no interrater estimate below .94. Perhaps a more realistic estimate of the probable reliability of this instrument in actual practice is given in Table 5, based on two observations by a team of three observers.

TABLE 4

Reliability Estimates for Five Separate Scales

Nine Teams of Three Raters
Observing Two Lessons (27 raters)

TYPE OF RELIABILTY	TOTAL SCALE 20 ITEMS	DOMAIN 2 4 ITEMS	DOMAIN 3 9 ITEMS	DOMAIN 4 4 ITEMS	DOMAIN 5 3 ITEMS
DISCRIMINANT	.91	.60	.89	.91	.63
STABILITY	.92	.61	.90	.94	.63
INTERRATER	.98	.88	.99	.98	.94

TABLE 5

Reliability Estimates for Five Separate Scales

One Team of Three Raters
Observing Two Lessons

TYPE OF RELIABILTY	TOTAL SCALE 20 ITEMS	DOMAIN 2 4 ITEMS	DOMAIN 3 9 ITEMS	DOMAIN 4 4 ITEMS	DOMAIN 5 3 ITEMS
DISCRIMINANT	.79	.31	.80	.81	.42
STABILITY	.86	.37	.85	.88	.42
INTERRATER	.85	.45	.89	.85	.63

Table 5 presents the reliability estimates for positive indicators resulting from a single team of three observers observing nine teachers each teaching two lessons. The results indicate a relatively high reliability for the total scale (20 items), as well as for Domains 3 (9 items) and 4 (4 items). Domain 2 (4 items) exhibits moderate reliability, however Domain 5 (3 items) estimates are below generally accepted levels for reliability. The lower reliability estimates for Domains 2 and 5 probably result from the small number of items in these subscales. In addition, Domain 5, Communication, appears to overlap all other domains, thus reducing its independence and resulting in increased ambiguity.

The high estimates for Domains 3, 4 and the Total scale suggest that they are appropriate for classroom application. The moderate estimates for Domain 2 suggest caution in its use as a subscale for evaluation. Domain 5 should probably not be used as a specific subscale based on these results.

Reliability of Ineffective Scales. The following table shows the output for the negative scores (ineffective indicators) from the instrument. Reliability estimates for the ineffective indicators are consistently lower than for the effective scores, and all ineffective scales exhibit questionable levels of reliability.

TABLE 6

Reliability Estimates for Ineffective Scales

One Team of Three Observers
Observing Two Lessons

TYPE OF RELIABILITY	TOTAL SCALE 20 ITEM	DOMAIN 2 2 ITEMS	DOMAIN 3 11 ITEMS	DOMAIN 4 4 ITEMS	DOMAIN 5 3 ITEMS
DISCRIMINANT	.64	.37	.40	.66	.58
STABILITY	.82	.67	.72	.77	.71
INTER-TEAM	.73	.55	.49	.69	.64

Possible reasons for the lower estimates from ineffective indicators include:

1. Far fewer instances of ineffective behaviors occurred in the study than of effective behaviors,

2. Ineffective indicators appear more diffuse (less clearly definable) than effective indicators, resulting in greater coding ambiguity, and

3. Several ineffective items require the observers to code "missing" behaviors (e.g. circulates inadequately, delays, etc.) Observers appear to code what the teacher **does** more accurately than what the teacher **does not do**.

Effects of Multiple Observers. The level of reliability resulting from using larger or smaller teams of observers was investigated by computing separate reliability estimates for randomly selected teams of observers. One team contained three observers, one contained two observers, and one team consisted of a single observer.

TABLE 7

Number of Observer's Effects on Reliability Estimates

One Team of Three Observers, Two Lessons

NUMBER OF OBSERVERS	TOTAL SCALE 20 ITEMS	DOMAIN 2 4 ITEMS	DOMAIN 3 9 ITEMS	DOMAIN 4 4 ITEMS	DOMAIN 5 3 ITEMS
			DISCRIMINANT		
THREE	.79	.61	.80	.81	.42
TWO	.75	.49	.76	.77	.38
ONE	.55	.25	.58	.64	.15
			STABILITY		
THREE	.86	.64	.85	.88	.42
TWO	.81	.55	.80	.85	.54
ONE	.70	.26	.77	.72	.37
			INTER TEAM		
THREE	.85	.66	.87	.85	.63
TWO	.82	.65	.85	.81	.54
ONE	.64	.50	.67	.71	.30

Table 7 indicates there is little difference between the reliability estimates if at least two observations are averaged to create scores. However, there is a considerable loss of reliability when only a single observer is used. These results suggest the use of at least two observers for evaluations.

Effects of Number of Visits per Rater Team.

TABLE 8

Reliability Estimates for Three Visits by One Team

INDICATORS	TOTAL SCALE 20 ITEMS	DOMAIN 2 4 ITEMS	DOMAIN 3 9 ITEMS	DOMAIN 4 4 ITEMS	DOMAIN 5 3 ITEMS
	==========DISCRIMINANT==========				
POSITIVE	.83	.39	.84	.85	.52
NEGATIVE	.68	.41	44	.71	.51
	========== ======STABILITY==========				
POSITIVE	.90	.4	.90	.92	.52
NEGATIVE	.87	.76	.79	.83	.48
	========== INTER RATER==========				
POSITIVE	.87	.51	.90	.88	.69
NEGATIVE	.75	.55	.51	.73	.54

Table 8 shows the effects of increasing the number of visits by each team. The reliability coefficients do increase over those obtained for two visits, but not substantially. This indicates that two visits by one team of two observers is probably optimum, with little gain from an increased number of visits.

Domain and Subscale Independence. Theoretically, effective and ineffective indicators in a specific domain should not correlate highly with each other. In addition, for the entire instrument and all subscales, a low correlation is expected between negative and positive subscale scores. To test for the independence cf effective and ineffective subtest scores on the summative instrument, all intercorrelations between these domain scores were computed. The results are shown in Table 9.

TABLE 9

Relationship of Effective to Ineffective Domain Scores

POSITIVE DOMAINS	NEGATIVE DOMAINS DOMAIN 2	DOMAIN 3	DOMAIN 4	DOMAIN 5
DOMAIN 2	-.111	.079	.189	.232
DOMAIN 3	.339	.107	.127	.398
DOMAIN 4	.359	.278	-.144	.140
DOMAIN 5	.218	.187	.003	.287

Most of the correlations are very low, only negative Domain 2 with positive Domains 3 and 4, (R=.33) and negative Domain 5 with positive Domains 3 and 5, (R=.39,.29) show any significant relationship, and then of a moderate degree. Only Domain 5 shows a significant positive correlation between effective and ineffective domain scores, and this is not surprising since Domain 5 (Communication), is involved in all teacher behaviors. Thus one may conclude that the positive and negative indicators are relatively independent.

Intercorrelations Among Positive Domain Scores.
As Table 10 shows, all domain scores correlate positively with the Total score, and indicate at least a moderate positive relationship to each other. Three relationships appear of particular interest: the Domain 3 score correlates highly with the Total score (r=.93), there is a negative correlation between Domain 2 and 4, and Domain 5 (Communication) correlates highly with all other scores except Domain 2.

Thus it appears from the moderate positive intercorrelations among these scales, that the domain scores are moderately related, but not identical. This strengthens the case for Domain 3, consisting as it does of 45% (9) of the total items on the instrument, dominates the relationships among domains, and total scores. This indicates that items in Domain 3 may be the best discriminators across teachers.

TABLE 10

Intercorrelations Among Positive Domain Scores

	TOTAL	DOM2	DOM3	DOM4
DOM2	.359	1.000		
DOM3	.932	.268	1.000	
DOM4	.543	-.285	.339	1.000
DOM5	.789	.167	.635	.500

Item Analysis. The ability of individual items to discriminate between "high scoring" teachers and "low scoring" teachers was tested in the following fashion:

1. Two groups (HI and LOW, were created to determine each item's ability to discriminate between teachers receiving high scores on the scale and those receiving low scores:

   a) Group 1 - consisted of the top six tapes on the summed total score.

   b) Group 2 - consisted of the bottom six tapes on the summed total score.

      i) By conducting t-tests between groups 1 and 2, we are able to estimate an item's ability to discriminate between effective and ineffective teachers as defined by this scale.

2. Since an item should discriminate between high and low scoring teachers, yet not discriminate between the same teacher on two separate occasions, the following pair of groups were created:

   a) Group A - consisted of each teacher's first lesson, and

b) Group B - consisted of each teacher's second lesson.

i) By conducting t-tests between groups A and B, we are able to determine whether an item fallaciously discriminates the same teacher from herself on two separate occasions.

Ideally, an item will indicate significant differences between groups 1 and 2, and no significant differences between groups A and B.
As Table 11 shows, most of the items discriminate between HI and LOW scoring teachers, yet do not discriminate between the same teacher on two different lessons. Only items 19 and 20 (Domain 2) appear to erroneously discriminate between two lessons by the same teacher. Reasons for this effect are as yet unknown, however, these items are somewhat different from the others as noted on page 6.

These results again support the superior discrimination of items in Domain 3, as all nine items attained a t-value of 6.41 or greater. Only five other items, two in Domain 2, achieved this level.

TABLE 11

Individual Item's Ability to Discriminate Among Teachers

ITEMS	T - VALUE BETWEEN GRPS 1 AND 2 HI vs LOW GROUPS	T - VALUE BETWEEN GRPS A AND B SAME TEACHER TWO LESSONS
1	1.00 *	.23
2	10.85	.16
3	7.19	1.39
4	10.05	3.27 **
5	13.21	1.02
6	10.48	.75
7	6.79	.41
8	7.15	3.60 **
9	8.19	.28
10	6.41	.12
11	6.71	1.26
12	1.00 *	.14
13	4.25	1.03
14	1.98 *	1.14
15	3.18	.62
16	1.00 *	1.64
17	10.18	.75
18	6.59	.54
19	6.77	7.93 **
20	7.40	8.09 **

* t value non significant (p<.01)
** t value significant (p<.01)

Internal Consistency Reliability Estimates. Although the scales developed from this instrument are not designed to be homogeneous, internal consistency often influences later research conducted using a specific instrument. For this reason, Coefficient Alpha estimates of internal consistency (Cronbach, 1951) were computed for all five positive (effective) scales for both raw and normalized data using the SPSS Reliability subprogram (release #9). Table 12 shows that the normalized items produce a higher estimate of internal consistency than do raw scores. The total score estimate (.69) is encouraging for a scale of this type which contains several apparantly independent subscales. This, in concert with the results of Item Discrimination indicates that the scale will probably correlate with other reliable measures dealing with teacher behavior, and be capable of differentiating between different groups.

TABLE 12

Internal Consistency Estimates for Five Scales

SCORE TYPE	TOTAL 20	DOM2 4	DOM3 9	DOM4 4	DOM5 3
RAW SCORES	.53	.44	.49	.45	.20
NORMALIZED	.69	.40	.63	.58	.37

Comparison of Repeated and Independent Estimates. In addition to the independent three way ANOVA conducted on the various scale and subscale scores for the summative instrument, a slightly different and generally more conservative estimate of reliability was conducted using a three way repeated measures ANOVA (Medley 1982). Table 13 compares the obtained estimates based on both independent and repeated measures ANOVAs. These estimates are for nine teams of three observers each observing two lessons.

As expected, the independent estimates are higher than the repeated estimates, for example Discriminant r = .91 vs. r = .76; however, the repeated estimates are quite high for this type of scale, providing further support for the reliability of the scales contained in this summative instrument.

TABLE 13

Repeated Measures vs Independent ANOVA Reliabiilty Estimates

Nine Teams of Three Observers
Two Lessons

SUBSCALE	Discriminant		Stability		Inter-team	
	Indep.	Repeat	Indep.	Repeat	Indep.	Repeat
Total (20)	.91	.76	.92	.85	.98	.82
Dom 2 (4)	.44	.30	.45	.34	.95	.66
Dom 3 (9)	.88	.80	.88	.81	98	.90
Dom 4 (4)	.91	.87	.94	.87	.98	.91
Dom 5 (3)	.63	.36	.63	.48	.94	.54

Factor Analysis. In an attempt to verify the structure of domain indicators, two separate factor analyses were conducted on the reliability study data, and compared to prior analysis conducted on training films using a preliminary form of the summative instrument.

Since the total sample of observations for the beginning teacher evaluation reliability study consisted of 768 observations (9 teachers, 2 lessons, 42 observers), and there were only 20 items, (only the positive items are included due to the inconsistency of negative indicators), it was possible to divide the total sample into two separate subsamples thereby allowing for an internal cross validation. One group was created from the first lesson for each teacher, and a second from the second lesson for each teacher.

Six factors were rotated during the analysis. Although the factors exhibited intransigence (were consistent from one sample to the other), no single factor appeared to account for a large amount of the variance (factor 1 - 15% was the greatest). This suggests that the factors are consistent, and located within the general domain structure of the sum-

mative instrument. Tables 14 and 15 provide a simplified depiction of the obtained factor loadings:

TABLE 14

Loadings for Factors One, Two and Three

ITEM #	DOM #	FACTOR #1 GRP1	FACTOR #1 GRP2	FACTOR #2 GRP1	FACTOR #2 GRP2	FACTOR #3 GRP1	FACTOR #3 GRP2
3	3	.44	.22	.21	.41		
4	3	.48	.32				
5	3	.82	.78				
6	3	.61	.75				
7	3	.80	.74				
2	3			.62	.62		
8	3			.66	---		
9	3			.64	.70		
10	3			.41	.70		
11	2-3			.33	.42		
17	5			.42	.31		
12	4					.84	.75
13	4					.80	.70
14	4					.46	.25
15	4					---	.51
16	5					.42	.60
20	2					.45	---

Factors 1 through 3 are more consistent across groups, more consistent with the domain structures, and more heavily weighted on important items than are factors 4 through 6. All of the factors examined tend to locate within a specific domain across both samples, and follow very closely the results obtained from a preliminary version of the instrument.

1. Factor 1 appears to load primarily on items: #5 - asks single, factual questions; #6 - asks questions requiring analysis or reasoning; and #7 - recognizes response, amplifies, gives corrective feedback. Factor 1 appears to be a questioning and response factor.

TABLE 15

Loadings for Factors Four, Five and Six

ITEM #	DOM #	FACTOR #4 GRP1	FACTOR #4 GRP2	FACTOR #5 GRP1	FACTOR #5 GRP2	FACTOR #6 GRP1	FACTOR #6 GRP2
1	2	.71	.75				
4	3	.51	.49				
16	5	.30	.30				
20	2	.35	.49				
11	2			.65	---		
14	4			---	.48		
15	5			.50	---		
19	2			.64	.81		
8	3					---	.61
14	4					.4.	.28
16	5	.30	.30			.40	---
17	5					.44	.50
18	5					.70	.75

2. Factor 2 appears to load primarily on items: #2 - Handles materials in an orderly fashion, #9 - provides for practice, and #10 - gives directions, assigns, checks comprehension, etc. Thus factor 2 appears to be an active interaction factor.

3. Factor 3 consists almost exclusively of elements from Domain 4; the various forms for presentation of subject matter.

4. Factors 4, 5 and 6 cross domains and appear more ambiguous than factors 1 thru 3. They tend to locate within Domains 2 and 5, which themselves are more diffuse than Domains 3 and 4, at least as measured by this instrument.

   a) Factor 4 appears to deal with timing and momentum, loading primarily on items #1 - Begins Instruction Promptly, #4 - Conducts Beginning, Ending Review, and #20 - Maintains Instructional Momentum.

   b) Factor 5 appears to relate to physical presence and active involvement, loading most heavily on

items #11 - Circulates and Assists Students, and #19 - Stops Misconduct (analysis conducted on the preliminary version of the instrument included only these two items in a factor, and in these analyses, they obtained the heaviest weights).

c) Factor 6 appears to be an enthusiam factor, loading as it does on items #17 - Expresses Enthusiasm, and #18 - Uses Body Behaviors That Show Interest.

These analyses indicate that the factors generated by this study tend to support the overall domain structure of the summative instrument. Since communication (Domain 5) is involved in all teacher behavior, it is not suprising to find it overlapping with factors 3, 4 and 6 in this analysis.

Limitations

Because the study was designed to assure that each rater observed the same teacher on the same two occasions, it is not possible to estimate precisely the amount of variation that may result from two or more raters observing a subject on different occasions. It is easier to generalize to the situation in which two raters observe a teacher simultaneously on two different occasions (four total observations).

In addition, by observing video tapes in a controlled situation, distractions that are almost certain to interfere with actual classroom observations were not present.

At this point in time, there is no effective way to use the instrument for evaluation purposes, as no information exists for comparison to general teacher patterns of behavior. It may be used in its current level for identification of problem areas and possible suggestions for remediation.

Conclusions

The results of this study tend to support the following conclusions:

1. The summative observation instrument appears sufficiently reliable (.80 or above) for total scores and major subscales.

   a) Subscales for Domains 2 and 5, however, should be used with caution.

2. A minimum of two observers should be involved in any evaluation using this instrument.

3. Positive indicators appear to be more reliable than negative indicators.

4. The structure identified within the Domains of teacher behavior appears to be supported by Factor Analytic results.

5. Domain 5 (Communication: Verbal and Nonverbal) appears to overlap with the other three domains in every test applied to the data.

6. Most items appear to discriminate between high scoring and low scoring teachers, and not to discriminate between the same teacher. This result indicates a degree of relative homogeneity for items in the total instrument. Internal consistency results also support this.

7. The effective and ineffective domains appear to be relatively independent, with the exception of Domain 5, which overlaps all other domains.

General Recommendations

This reliability study has produced positive results and the summative instrument appears sufficiently reliable to produce consistent scores for teacher behavior when at least two observers are used to gather data. At this point in the investigation of the instrument, we may expect these results to generalize to various teachers, various raters, and the same teacher in more than one situation. Although reliability estimates will probably be lower for actual classroom application, they appear to be high enough that a small to moderate loss in the "Real World" environment will still yield useful results.

Specific Recommendations

Due to the lower reliability estimates resulting from the ineffective domains, we recommend that the highly reliable effective domain scores be used to identify general areas for remediation; followed by investigation of specific ineffective observations as indicators of particular teacher practices that may require remediation.

We recommend that standard normalized scores based on a large sample of teachers be developed to provide a basis for the evaluation of a specific teacher.

We recommend that at least two separate observations by at least two observers be used as a basis for an individuals scores in evaluation. The mean scores from two separate observers will be compared to standard normalized scores.

We recommend that this instrument ultimately be used as a part of the evaluation and remediation process for individual teachers.

Future Research

At the present time, the following topics are being investigated:

1. INSTRUMENT VALIDATION,
2. RATER (OBSERVER) CERTIFICATION,
3. NORMING, and
4. FRAME FACTORS.

Validation

In the future, research will be conducted to support validity by relating scores on this instrument with both ratings of teacher effectiveness, and achievement scores of students.

Rater Certificaton

Prior to an individuals use of this instrument they must receive sufficient training to accurately and reliably code indicators. This will probably be measured in the following fashion: students will observe two tapes of the same teacher teaching the same lesson (or a very similar lesson). Their observations will be compared with an average resulting from several trained raters (all tapes used in this study have at least 27 observer records), and a correlation will be computed. In addition, a correlation will be computed between the individual's first observation and his/her second observation. In this way, two measures of a prospective rater's skills will be obtained (Frick and Semmel 1978):

1. ACCURACY - the relationship to a pre-established master score, and
2. TEST-RETEST - the relationship between observations of two almost identical situations at two different times by the same rater.

Norming

Prior to the use of this instrument for evaluation it will be necessary to determine behavior patterns of currently practicing teachers. This data must be viewed in light of at least three factors:

1. Frame Factors - socio-demographic factors that appear to affect student outcomes,

2. Behavior Patterns - specific behaviors will tend to associate with certain other behaviors. Absolute numbers of behaviors may not be as important as the relationship among the specific behaviors.

3. Normalized (standardized) Scores - based perhaps upon averages for specific certified observers across all of their observations. It is assumed that some observers will be lenient, while others will be strict in their interpretation of specific indicators. While the use of at least two observers should control for this effect somewhat, even team scores will tend to be either strict or lenient, and this must also be considered in data interpretation. Scores normalized across rater teams should provide the best source of information concerning an individual teacher's use of effective behaviors.

Frame Factors

Presently, several variables that have shown historical relationships with student outcome variables are being obtained along with teacher scores on the instrument. These include:

1. Academic status of students
2. Socio-economic status of students
3. Non-native language speakers
4. Exceptionalities of students
5. Class size
6. Sex of students
7. Classroom conditions
8. Instructional material conditions
9. Teacher variables
   - 9.1 Experience
   - 9.2 Education
   - 9.3 Tenure
   - 9.4 Area of certification
10. Subject (math, English, etc.)
11. Grade level (elementary, middle, high school)

Tests will be conducted to determine which if any of these factors significantly influence teacher behavior as measured by the summative instrument. Regression analysis will be used to identify the most "important" factors, and may result in differential norms for different factor combinations.

BIBLIOGRAPHY

Brennan, Robert L., and Michael T. Kane. An index of dependability for mastery tests. J. of Ed. Measurement 1977, 14:3, 277-289

Cronbach, L.J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16:3, 297-332

Cronbach, L.J., Clesser, G.C., Nanda, J., and N. Rajaratnam. The Dependability of Behaviorial Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley, 1972.

Frick, Ted, and Melvyn I. Semmel. Observer agreement and reliabilities of classroom observational measures. Review of Ed. Rsch. 1978, 48:1, 157-184.

Guilford, J.P., and B. Fruckter. Fundamental Statistics in Psychology and Education. New York: Mcgraw-Hill, 1973

Lahey, Mary Anne, Ronald G. Downey, and Frank E. Saal. Intraclass correlatons: there's more there than meets the eye. Psychological Bulletin 1983, 93:3 585-595

Lindeman, Richard H., P. F. Merenda, and R.Z. Gold. Introduction to Bivariate and Multivariate Analysis. Glenview, Ill: Scott, Foresman and Co., 1980

Medley, Donald M., and Harold E. Mitzel. Measuring classroom behavior by systematic observation. In N.L. Gage (Ed.), Handbook of Research on Teaching. Chicago,Ill.: Rand-McNally, 1963, 247-328.

Medley, Donald M., H. Coker, J. Coker, J. L. Lorentz, R. S. Soar, and Robert L. Spaulding. Assessing teacher performance from observed competency indicators defined by classroom teachers. J. of Ed. Research, 1981, 74:4, 197-216.

Medley, Donald M., Personal correspondence, 1982.

Mitchell, Sandra K. Interobserver agreement, reliability, and generalizability of data collected in observational studies. Psychological Bulletin 1979, 86:2, 376-390

Nunnally, Jum C. Psychometric Theory. New York: McGraw-Hill, 1978.

Shrout, Patrick E., and Joseph L. Fleiss. Intraclass correlations: uses in assessing rater reliability. Psycholocial Bulletin, 1979, 86:2, 420-428.

Soar, Bob. Personal correspondence, 1982.

Appendix A

Knowledge Base and the Florida Teacher Competencies

The development of instruments for measuring classroom performance of teachers requires that a body of information about teaching be assembled. Such information can be derived from two sources: first, the consensus of opinion of informed persons such as teachers and pedagogical instructors about the knowledge and skills believed to be necessary for effective classroom performance: and second, process-product and experimental research on teacher effectiveness.

The Florida Competencies were derived from consensus of opinion among informed school people. While these competencies are useful for general purposes, the original research team chose to turn to research literature as the source of a knowledge base for instrument development. This approach was selected for the following reasons:

1. The knowledge and skills derived from research literature, as the knowledge base for evaluating teachers, is easier to defend if contested.

2. The research studies indicate precisely what teacher behaviors are positively associated with either student achievement or student conduct or both.

3. The research literature enables one to cite specific evidence in support of particular teacher performance. The language of research studies is precise and thus allow little chance for misinterpretation.

4. The research findings relevant to teacher effectiveness provide grounds for an examination of the Florida competencies.

Appendix β

Indicators of the Summative Instrument

Indicators of teacher behavior are divided into two types, EFFECTIVE (positive) and INEFFECTIVE (negative). The following descriptors are used on the summative instrument to specify essential characteristics of behaviors belonging to a particular domain:

TABLE 16

Summative Instrument Descriptors - Domain 2

POSITIVE INDICATORS		NEGATIVE INDICATORS	
ITEM #	DESCRIPTORS	ITEM #	DESCRIPTORS
1	BEGINS INSTRUCTION PROMPTLY	1	DELAYS
11	CIRCULATES AND ASSIST STUDENTS	10	REMAINS AT DISK/CIRCULATES INADEQUATELY
19	STOPS MISCONDUCT	19	DELAYS DESIST/DOESN'T STOP MISCONDUCT/DESISTS PUNITIVELY
20	MAINTAINS INSTRUCTIONAL MOMENTUM	20	LOSES MOMENTUM/FRAGMENTS NON ACADEMIC DIRECTIONS,

TABLE 17

Summative Instrument Descriptors - Domain 3

POSITIVE INDICATORS		NEGATIVE INDICATORS	
ITEM	DESCRIPTORS	ITEM	DESCRIPTORS
2	HANDLES MATERIAL IN AN ORDERLY MANNER	2	DOES NOT ORGANIZE OR HANDLE MATERIALS SYSTEMATICALLY
3	ORIENTS STUDENTS TO CLASSWORK/MAINTAINS ACADEMIC FOCUS	3	ALLOWS TALK/ACTIVITY UNRELATED TO SUBJECT
4	CONDUCTS BEGINNING/ ENDING REVIEW		
5	ASKS SINGLE/FACTUAL QUESTIONS	4	POSES MULTIPLE QUESTIONS ASKED AS ONE/UNISON RESPONSE
6	ASKS QUESTIONS REQUIRING ANALYSIS OR REASONING	5	POSES NON-ACADEMIC QUESTION/ NON-ACADEMIC PROCEDURAL QUESTIONS
7	RECOGNIZES RESPONSE/ AMPLIFIES/GIVES CORRECTIVE FEEDBACK	6	IGNORES STUDENT OR RESPONSE/ EXPRESES SARCASM, DISGUST OR HARSHNESS
8	GIVES SPECIFIC ACADEMIC PRAISE	7	USES GENERAL, NON SPECIFIC PRAISE
9	PROVIDES FOR PRACTICE	8	EXTENDS DISCOURSE, CHANGES TOPIC WITH NO PRACTICE
10	GIVES DIRECTIONS/ ASSIGNS/CHECKS COMPREHENSION OF HOMEWORK SEATWORK ASSIGNMENT/ GIVES FEEDBACK	9	GIVES INADEQUATE DIRECTIONS/ NO HOMEWORK/NO FEEDBACK

TABLE 18

Summative Instrument Descriptors - Domain 4

POSITIVE INDICATORS		NEGATIVE INDICATORS	
ITEM #	DESCRIPTORS	ITEM #	DESCRIPTORS
12	TREATS CONCEPT/DEFINITION/ATTRITUBES/EXAMPLES/NON-EXAMPLES	11	GIVES DEFINITION OR EXAMPLES ONLY
13	DISCUSSES CAUSE-EFFECT EFFECT/USES LINKING WORDS/APPLIES LAW OR PRINCIPLE	12	DISCUSSES EITHER CAUSE OR EFFECT ONLY/USES NO LINKING WORDS
14	STATES AND APPLIES ACADEMIC RULE	13	DOES NOT STATE OR DOES NOT APPLY ACADEMIC RULE
15	DEVELOPS CRITERIA AND EVIDENCE FOR VALUE JUDGEMENT	14	STATES VALUE JUDGEMENT WITH NO CRITERIA OR EVIDENCE

TABLE 20

Summative Instrument Descriptors - Domain 5

POSITIVE INDICATORS		NEGATIVE INDICATORS	
ITEM #	DESCRIPTORS	ITEM #	DESCRIPTORS
16	EMPHASIZES IMPORTANT POINTS		
17	EXPRESES ENTHUSIAM VERBALLY/CHALLENGES STUDENT		
		15	USES VAGUE/ SCRAMBLED DISCOURSE
		16	SES LOUD-GRATING, HIGH PITCHED, MONOTONE, INAUDIBLE TALK
18	USES BODY BEHAVIOR THAT SHOWS INTEREST/ SMILES, GESTURES	18	FROWNS, DEADPAN OR LETHARGIC/ OVERDWELLS

Appendix C

Computation of Reliability Estimates

The following tables indicate the use of Sums of Squares generated by a Three Way ANOVA on the Total Score (20 items) to produce three reliability outputs: (1) discriminant, (2) stability, and (3) interrater. For the development of these formulations see (Medley and Mitzel, 1963, Shrout, 1979).

TABLE 21

Sources of Variance for Reliability Estimates

SOURCE	N	DF	SS'S	MEAN SQ	LETTER
TEACHER	9	8	3435.69118306	429.4610	a
LESSON	2	1	18.28686575	18.2869	b
RATERS	9	8	213.04918227	26.6310	c
TEACH*LESS		8	273.17772720	34.1472	d
TEACH*RATE		64	540.44689727	8.4444	e
LESS*RATE		8	83.17933286	10.3974	f
RATE*LESS*TEACH		64	277.97843948	4.3434	g

Using the mean squares shown above, the independent ANOVA estimates are produced in the following fashion:

TABLE 22

Computations for Total Scale

ONE TEAM OF RATERS
TWO LESSONS
NINE TEACHERS

K = NUMBER OF LESSONS (SITUATIONS) IN ESTIMATE = 2
N = NUMBER OF RATER TEAMS IN SOURCE STUDY = 9
N1 = NUMBER OF RATER TEAMS IN THIS ESTIMATE = 1

TRUE SCORE = (K*N1)(A - D - E + G)/2N
= 2(429.460-34.147-8.444+4.343)/9
= 43.468

ERROR SCORE = (N1*(D - G))/N + K(E - G)/2 + G
= (34.147-4.343)/9 + 2(8.444-4.343)/2 + 4.343
= 11.756

ERROR(S) STABILITY = (N1*(D-G))/N + G
= (34.147-4.343)/9
= 7.655

ERROR(R) RATER = K(e - g)/2 + g
= 2(8.444-4.343)/2 + 4.343
= 8.444

OBTAINED SCORE = TRUE + ERROR
= 43.468 + 11.756
= 55.224

r (Discrimant) = 1 - ERROR/OBTAINED
= 1 - 11.756/55.224
= .787

p (Stability) = 1 - ERROR(S)/OBTAINED
= 1 - 7.655/55.224
= .861

q (Interrater) = 1 - ERROR(R)/OBTAINED
= 1 - 8.444/55.224
= .847
